第11卷 第20期 2011年7月
1671—1815(2011)20-4730-05

科 学 技 术 与 工 程
Science Technology and Engineering

Vol. 11 No. 20 July 2011

医药卫生

机器学习技术在胸癌诊断中的应用

李 蓉 孙 媛

(北京物资学院信息学院,北京 101149)

摘 要 为了提高胸癌诊断的识别精度,提出了应用机器学习方法建立胸癌诊断模型。其中描述细胞特征的参量作为模型的输入,细胞的类别对应模型的输出。选取三种机器学习方法作为建立模型的训练算法,分别为反向传播(Back Propagation, BP)神经网络、学习矢量量化网络(Learning Vector Quantity, LVQ)和支持向量机(Support Vector Machine, SVM)。仿真结果显示三种机器学习方法所见的诊断模型均具有较高的识别率(BP:97.28%, LVQ: 98.06%, SVM: 98.45%),可作为有效地识别方法用于其他医学诊断研究。

关键词 神经网络 特征参量 支持向量 权值 学习矢量

中图法分类号 R734.3 TP183; **文献标志码** A

尽管近年来医学研究取得了较大发展,胸癌仍是威胁人类健康的主要疾病之一。在胸癌研究中,诊断和预测是两个主要研究领域,本文只涉及诊断过程。胸癌诊断通常采用活细胞组织检查来实现,诊断中存在很多和肿瘤相关的特征,特征相互作用使得诊断过程依赖医生的经验。

为了解决传统诊断方法中的主观因素,需要提出自动诊断识别技术。传统的人工智能和神经网络技术已被应用并取得了较好的效果[1,2]。为了进一步提高胸癌诊断模型的识别能力,本文尝试采用三种不同的机器学习方法建立胸癌诊断模型。所采用的机器学习方法分别为 BP 神经网络、学习矢量量化和支持向量机。其中 BP 神经网络是一种常用的神经网络模型,学习矢量量化是一种基于竞争学习规则的新型神经网络,支持向量机是基于统计学习理论的机器学习方法。

建模过程采用求解模式识别问题的方法,针对应用问题提取特征参量,选择分类方法建模构造分类器,并对模型进行诊断检测。本文分别采用上述三种机器学习方法构建了诊断模型,验证这些方法在胸癌识别中的有效性。

1 诊断原理

胸癌诊断是根据描述细胞形态和活性检查结果等参量来判断该细胞的类别,本质上一个多元识别问题。应用机器学习建立诊断模型首先基于将诊断过程理解为一个模式识别问题。求解模式识别问题的一般步骤是首先选取特征参量和对参量预处理,第二步采用机器学习方法设计分类器,也是算法的训练过程,最后将特征参量代入建好的模型分类决策[3]。

诊断过程遵循求解模式识别问题的步骤,首先选择描述细胞的特征参量和最终类别建立训练集,训练集的样本是以细胞为单位的输入和输出数据。其中特征参量作为样本输入,如果该细胞经医学诊断为恶性,我们把此样本归为正例样本,样本的输出为 +1。否则此样本为反例样本,输出为 -1。第二步把建立好的训练集带入机器学习的训练算法创建诊断模型。最后将待识别的细胞得属性参量

2011年3月24日收到 国家自然科学基金重点项目、国家自然科学基金项目(10973020)、北京市属高等学校人才强教计划资助项目(PHR200906210)、北京市教育委员会科研基地建设项目(WYJD200902)、"十一五"国家科技支撑计划重点项目课题、区域性国际物流综合服务系统与应用示范"北京市哲学社会科学规划项目(09BaJG258)资助

代入诊断模型就可以诊断其类别。

2　诊断模型描述

本次建模的数据来源于美国 Wisconsin 大学采集的胸癌数据,由 http://www. ailab. si/orange/doc/datasets/breast-cancer-wisconsin-cont. tab 下载。数据集中共有 683 个样本数据,每个样本数据包括 9 个描述细胞特征的参量和一个类别。具体的描述如下表所示:

表 1　Wisconsin 胸癌数据集描述

属性编号	属性名称	值
1	肿瘤厚度	1 – 10
2	细胞尺寸	1 – 10
3	细胞形状	1 – 10
4	末端纤维黏连度	1 – 10
5	上皮单细胞尺寸	1 – 10
6	裸细胞核	1 – 10
7	染色质	1 – 10
8	正常细胞核	1 – 10
9	间接核分裂	1 – 10
10	类别	0,1

下面分别介绍这三种机器学习方法及其在胸癌诊断中的应用。

2.1　基于 BP 神经网络的诊断模型

BP 网络的训练是采用反向传播算法的有教师学习方法[4]。整个学习过程分为正向传播和反向传播。当正向传播时,输入信息从输入层经隐层单元处理后传向输出层;另一个是反向传播过程,将误差信号沿原来的神经元连接通路返回,逐层计算梯度修改各层神经元连接的权值。直至算法收敛。具体的算法如下所述:

Step1　对学习步幅 ρ 选一个小的正数,对所有神经元的连接权值 $\{w_{ij}\}$ 赋以小的随机初值。

Step2　重复以下各步直到权的变化以及均方误差 ε 的变化足够小。

(1)取下一个训练样本 E 及其对应的正确输出 C。

(2)前向传播步骤:由输入层至输出层计算每个神经元的加权和 y_i 以及输出 $X_i = f(y_i)$。

(3)反向传播步骤:由输出层开始逐层计算输出层和隐层神经元的如下量:

$f'(y_i) = x_i(1 - x_i)$;

$$\delta_i = \begin{cases} (c_i - x_i)f'(Y_i), \text{如果 } x_i \text{ 是输出单元} \\ (\sum_{m>i} w_{mi}\delta_m)f'(Y_i), \text{如果 } x_i \text{ 不是输出单元。} \end{cases}$$

(4)更新权值

$W'_{ij} = W_{ij} + \rho\delta_i y_j$。

上述算法中 ρ 为训练步长,$0 < \rho < 1$。$y_i = \sum_j W_{ij}x_j$ 表示某一层的第 i 个神经元的输入加权和,x_j 是与其相连的下一层个神经元的活性,W_{ij} 表示下一层第 j 个神经元与其连接的权。算法采用了 Sigmoid 函数作为神经元的激发函数 $f(x) = \frac{1}{1 + e^{-x}}$。

应用 BP 网络诊断模型中,对细胞属性的描述参量作为模式识别问题的输入参量,我们的数据中共有 9 个属性,表示为向量 $x_i = (x_{i1}, x_{i2}, \cdots, x_{i9})$。细胞的类别作为问题的输出 $y \in \{1,0\}$:恶性细胞作为正例样本,标记为 1;良性细胞作为反例样本,标记为 0。诊断模型建立一个

三层的神经网络(图 1),包括输入层,隐层和输出层。输入层由 9 个神经元组成,对应用描述细胞的 9 个特征参量。隐层单元设为 4 个神经元,输出层由一个神经元组成。将一类对应于输出 1,另一类对应输出 0,识别时只要输出大于 0.5 则决策为第一类,否则决策为第二类。

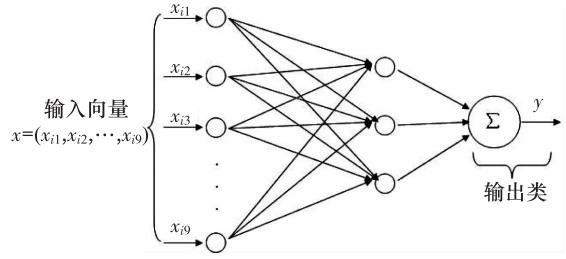

图 1　BP 网络诊断模型

2.2　基于 LVQ 的诊断模型

学习矢量量化是在矢量量化基础上对输入向

量分类的一种有监督学习技术[5]。LVQ 在形式上属于一种竞争学习的神经网络。由输入层和输出层组成,根据竞争学习规则,获胜单元是离输入单元最近的一个输出单元,只有连接输出单元的权值被修改。

我们用$\{x_j\}$ $(j=1,2,\cdots,N)$表示输入向量集,这里 x_j 代表第 j 的样本的描述细胞属性的向量。$\{w_j\}$ $(j=1,2,\cdots,m)$表示网络突触权值向量。我们用 C_{w_j}表示与权值向量 w_j 相关的分类,并且 C_{x_i}是网络输入向量 x_i 的类标签,对应 x_i 表示的细胞样本为恶性或者为良性。权值向量以下面方式来调整。

如果输入向量与权值向量同类,则权值向量 w_j 沿着输入向量 x_i 方向移动;如果类别不同,根据公式(2),权值向量 w_j 沿着远离 x_i 方向移动。具体的算法如下:

LVQ 算法:

步骤 1 初始化所有权值向量 $w_j(0)$,初始化学习率参数 $\mu(0)$,并且设置 $k=0$。

步骤 2 检查终止条件。如果失败,继续;如正确,退出。

步骤 3 对每个训练向量 x_j 执行步骤 4 与步骤 5。

步骤 4 决定权值向量标签 $(j=q)$,满足 $\min\limits_{\forall j}\|x_i-w_j(k)\|_2^2$。

步骤 5 更新权值向量 $w_q(k)$如下:

如果 $C_{w_q}=C_{x_i}$,则

$w_q(k+1)=w_q(k)+\mu(k)[x_i-w_q(k)]$;

如果 $C_{w_q}\neq C_{x_i}$,则

$w_q(k+1)=w_q(k)-\mu(k)[x_i-w_q(k)]$。

步骤 6 用 $k+1$ 代替 k,降低学习率参数,再回到步骤 2。

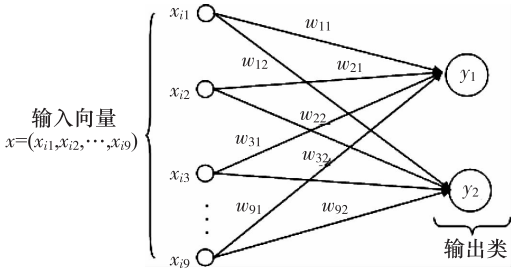

图 2 LVQ 网络诊断模型

2.3 基于 SVM 的诊断模型

SVM 是基于结构风险最小化(structural risk minimization, SRM)原则的机器学习方法,具有较好的推广能力[6]。SVM 采用核化技术,用内积核函数映射原始特征空间的样本 x 到一个高维特征空间的$\oint(x)$,在这个空间构造最优分类超平面,即寻求一个特征空间的平面不仅能把样本分开,还是分界面的间隔最大。求解最优分类超平面可以变换为一个不等式约束下的二次优化问题[7],其中(x_i,y_i)对应训练集中的输入输出样本对,即细胞属性及类别。

$$W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j k(x_i\cdot x_j) \quad (1)$$

约束条件为:

$$0\leqslant\alpha_i,i=1,\cdots,l\text{ 与 }\sum_{i=1}^{l}\alpha_i y_i=0 \quad (2)$$

式(2)中 α_i 不为零对应的样本就是支持向量。

训练结束后将带识别的细胞样本 x 代入下面的分类函数决策

$$f(x)=\operatorname{sgn}(\sum_i y_i\alpha_i(x_i,x)-b) \quad (3)$$

诊断建模中的 SVM 的网络结构同于图 1 的 BP 网络,不同之处是隐层单元为核函数,隐层单元的个数为训练后所求的支持向量的个数。

3 实验结果

实验中采用 Bootstrap 方法划分数据集,具体操作为对 n 个样本的数据集随机抽取 n 次,将取到的样本去掉重复作为训练样本,余下的样本作为测试样本。这种抽样方法所得到的训练样本集约占总数据集的 63.2% 。重复上述操作十次,取平均结果。采用该方法划分得到训练集 425 个样本,测试集 258 个样本。

将训练集样本代入上述三种机器学习方法建立各自的诊断模型。其中 BP 网络建模使用 MATLAB 神经网络工具箱[7],批梯度下降训练函数 traingd,误差精度设为 0.01,迭代次数 250。LVQ 算法是 WEKA(Waikato 知识分析环境)开发的软件

包,软件用 JAVA 语言编写,由 http://wekaclassalgos.sourceforge.net 网站下载。初始参考点的数目设为和正例样本数相同的个数,训练的迭代次数设为 10 000,其他参数保持默认设置。SVM 模型中核函数选择高斯核函数参数 g 取值 0.005,错误惩罚参数 $C=5$,实验结果如下表 1 所示。

在训练建立诊断模型后,对诊断模型进行了开放测试。测试集包括 258 个样本,其中正例样本(恶性细胞类)92 个,反例样本(良性细胞类)166 个。测试集代入建立的诊断模型,得到测试结果如表 2 至表 4 所示。

表 2　BP 模型的识别结果

诊断输出/%	期望输出/%		
	恶性细胞	良性细胞	全部
良性细胞	163 (98.19)	4 (1.55)	167
恶性细胞	3(1.55)	88 (95.65)	91
全部	166	92	258 (97.28)

表 3　LVQ 模型的识别结果

诊断输出/%	期望输出/%		
	恶性细胞	良性细胞	全部
良性细胞	163 (98.19)	2 (1.55)	165
恶性细胞	3 (1.55)	90 (97.83)	93
全部	166	92	258 (98.06)

表 4　SVM 模型的识别结果

诊断输出/%	期望输出/%		
	恶性细胞	良性细胞	全部
良性细胞	163 (98.19)	1(1.55)	164
恶性细胞	3 (1.55)	91 (98.91)	94
全部	166	92	258 (98.45)

由表 2 至表 4 可以看出三种机器学习方法建立的诊断模型具有较高的胸癌识别率和正确率。三种方法识别良性细胞能力相同,都为 163 个。对恶性细胞即胸癌的识别率有所不同。其中 SVM 算法的胸癌识别率最高(98.91%),其次是 LVQ 算法(97.83%)和 BP 算法(95.65%)。

原因是 SVM 算法的分类能力较强,对位于分界面附近的样本有较好的识别能力。LVQ 是基于参考点的方法,在很多应用中有稳定的分类性能。BP 算法采用了传统的反向传播算法,在诊断模型中的性能较好,在某些应用中会出现过学习和局部最小化现象。

4　结论

胸癌诊断是医学研究的一个主要领域,传统上凭借医生的经验诊断。本文将诊断过程形式化为模式识别问题,应用机器学习算法建立了诊断模型。模型的输入为细胞的特征参量,输出为细胞所属的类别。

机器学习方法选取为 BP 神经网路,LVQ 方法和 SVM 方法。开放测试结果显示三种方法所建的模型均具有较高的诊断能力。作为传统的神经网络方法,BP 网络在诊断问题中体现了较好的识别分类能力,可以作为一种简单可行的诊断方法。LVQ 算法也体现了较强的识别性能,算法运行速度快。正如在其他领域中的成功应用,SVM 方法在胸癌识别中展示了较好的识别能力。这三种机器学习方法可以用于将来的胸癌诊断研究和其他医学诊断建模。

参　考　文　献

1　Quinlan J R. Improved use of continuous attributes in C4.5. Journal of Artificial Intelligence Research,1996;4:77—90

2　Karabatak M, Cevdet M. An expert system for detection of breast cancer based on association rules and neural network. Expert Systems with Applications, 2009;36: 3465—3469

3　边肇祺,张学工. 模式识别. 北京: 清华大学出版社,2000

4　杨建刚. 人工神经网络实用教程,浙江:浙江大学出版社,2002

5　叶世伟. 神经计算原理. 北京: 机械工业出版社, 2007

6　Vapnik V N. The nature of statistical learning theory. NY: Springer Verlag, 1995

7　闻　新,周　露,李　翔. MATLAB 神经网络仿真与应用. 北京: 科学出版社,2003

(下转第 4739 页)

The Design of Recirculating Aquaculture Temperature Control System Based on PLC

WU Yan-xiang , HU Yong-mei , LIU Yu-qing
(College of Engineering , Shanghai Ocean University,Shanghai 201306,P. R. China)

[**Abstract**] Since study on the domestic aquaculture monitoring system was backward at present, the SIMATIC S7—200 CPU226 is used and adopted PID control algorithm to make real-time monitoring and controlling for the water level and temperature of closed circulating aquaculture system, thus it realized constant water level and temperature control in the culture pond. The system overall control scheme and principle are introduced in detail, and the system software and hardware structure design are given, the system has efficiently implemented system logic control, safety control, fault display and fault treatment.

[**Key words**] PLC PID algorithm closed circulating aquaculture water level temperature

(上接第 4733 页)

Breast Cancer Diagnosis Using Machine Learning Technique

LI Rong, SUN Yuan
(The Institution of Information, Beijing Wuzi University, Beijing 101149,P. R. China)

[**Abstract**] In order to improve the diagnosis accuracy, machine learning method was proposed to construct the breast cancer diagnosis model. The parameters of cell feature are the inputs of model and the class of diagnosed cell is the output. Three machine learning methods are chosen as training algorithm, including BP neural network, learning vector quantity network and support vector machine. Simulation results show that three methods have high identification ability (BP:97.28%, LVQ: 98.06%, SVM: 98.45%) and can be applied to other medicine research as effective method.

[**Key words**] neural network feature parameters support vector weight learning vector